图书情报工作
第 62 卷 第 20 期 2018 年 10 月

知乎平台用户影响力分析与关键意见领袖挖掘

■ 郭博[1] 许昊迪[2] 雷水旺[3]
[1] 珠海市魅族科技有限公司北京分公司 北京 100872 [2] 香港科技大学 香港 999077
[3] 徐州工业职业技术学院图书馆 徐州 221140

摘要：［目的/意义］随着互联网技术的快速发展，知乎平台逐渐成为一个热议社会公众话题以及分享知识、经验的载体。因此，分析知乎平台中关键用户的影响力和挖掘其中的关键意见领袖在研究社交网络信息传播途径的过程中起到非常重要的作用。［方法/过程］通过提出改进的 PageRank 算法和 HITS 算法，分别基于知乎用户社交网络、问答网络构建用户影响力挖掘模型，能够准确、客观地识别出其中的关键用户及意见领袖。［结果/结论］实验结果表明，提出的 PageRank 算法和 HITS 算法能够有效挖掘出知乎平台中具有较为突出特性的关键意见领袖，并且算法的收敛速度较快，具有可复用性和迁移性。通过对知乎平台用户数据集进行处理和有效分析，成功建立用户影响力和关键意见领袖挖掘模型；同时，在具体话题上进行验证。因此，可以推断该模型有巨大应用价值和商业化推广前景。

关键词：知乎 用户影响力 关键意见领袖 PageRank 算法 HITS 算法
分类号：TP393
DOI：10.13266/j.issn.0252-3116.2018.20.014

1 引言

随着社交网络的发展，微博、知乎等社交软件逐渐成为各行各业精英用来交流行业信息、专业知识以及生活经验的常用工具[1]。特别的是，知乎提供的问答平台不仅满足了用户之间基本的社交需求（相互关注及访问），也满足了在分享和问答的过程中建立个人威望以及自我实现的人性需求，因此吸引了大量的行业精英、专家、网络红人、社会名人等权威人士加入其中，并根据其相互之间的关系逐步形成一个具有多主体直接互动性、长期性以及近同时性的社交问答网络。

关键意见领袖是指在人际传播网络中为他人提供信息，同时也对他人施加影响的“活跃分子”[2]。他们既是信息传播过程中的权威起源者，也是主要扩散者，由此形成了社交网络中的信息级联传播。在社交网络的信息传播交流过程中，具有较强影响力的用户以及关键领袖在其中发挥的作用越来越明显。近年来的大量研究表明，关键意见领袖在网络知识传播[3]、网络口碑效应[4]以及网络交际[5]等社会现象中起到了非常重要的影响。随着知乎平台日益成为较为重要的知识分享问答社区，研究用户的影响力并识别出其中的关键意见领袖对于网络干预[6]、网络营销[7-9]、网络结构分析以及电子商务领域具有极大的意义。

在大多数识别关键意见领袖的文献中，使用社交网络分析的方法与其他方法相比具有较为明显的优势[10]。为了在知乎网络中综合性地识别关键用户，笔者利用知乎平台中用户之间的社交以及问答等互动行为信息，建立了一个基于知乎的多层次的综合性评价网络，如图 1 所示。在图 1 中，每一个圆圈结点代表一个用户，方块结点代表一个回答，而它们之间的连线代表其相互的社交关系以及问答关系。笔者利用其社交关系、问答关系以及其之间的联系，建立一个双层的知乎网络结构。

根据先前建立的双层知乎网络结构，首先使用层次分析法构建全方位的知乎用户评价指标，并建立用户影响力评价模型。使用用户影响力评价模型可以得到每一位知乎用户在层次分析指标上的影响力权重，随后将该结果与传统的 PageRank 算法和 HITS 算法模型进行结合，得到改进的综合性评价模型，用于挖掘知乎平台中的关键用户及意见领袖。该研究过程及其框架如图 2 所示：

作者简介：郭博（ORCID：0000-0002-4053-749X），数据专家，高级工程师，博士研究生，E-mail：guobo01@126.com；许昊迪（ORCID：0000-0001-6016-9084），硕士研究生；雷水旺（ORCID：0000-0002-7120-4460），馆员，硕士。
收稿日期：2018-03-28 修回日期：2018-07-19 本文起止页码：122-132 本文责任编辑：徐健

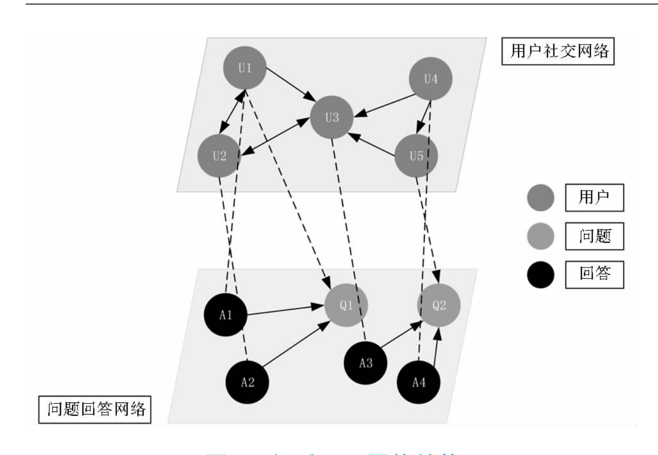


图 1　知乎双层网络结构

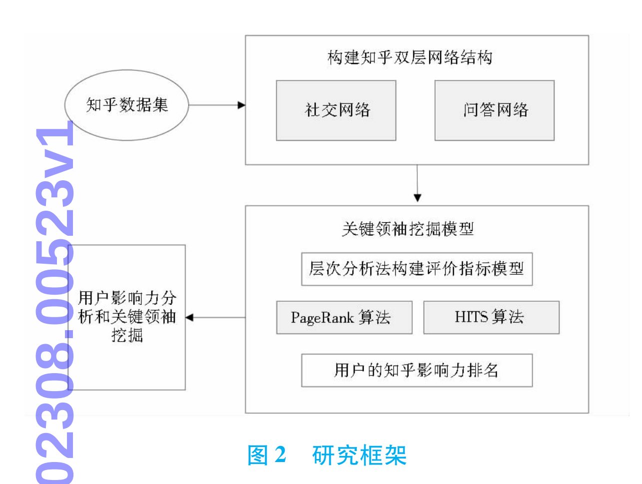


图 2　研究框架

基于对知乎群体和其社交网络特性上的分析和探索，笔者提出了一种基于改进的 PageRank 和 HITS 算法模型用于挖掘知乎平台上的关键用户以及分析其影响力。相对于现有的研究成果，该模型具有以下的贡献：

首先，该模型结合了知乎平台用户之间的社交网络和问答网络关系来评估一个用户是否是关键用户。虽然用户的影响力和权威度可以通过社交网络体现出来，而其信息的传播能力展现在活跃的问答行为中。因此，将这两个网络层的信息相结合，有助于综合、客观地挖掘出知乎平台中的关键用户和评估其在整个平台中的影响力。

其次，用于分析用户影响力和挖掘关键意见领袖的传统 PageRank 算法和 HITS 算法都更多地考虑到用户或者问答之间的链接数目，而很少评估用户的可信程度以及问答的质量，因此很难客观地评价用户的真正影响力[11]。笔者根据层次分析法确定用户的综合评价模型用于改进传统的 PageRank 算法和 HITS 算法模型，通过加入对用户活跃度、可信度和影响力的评估，并且加权两种算法的结果进行排名，能够较大程度上改进原有模型在初始平分权重带来的限制，进一步提升模型的准确性和有效性。

最终，笔者提出了一种可复用的、客观的关键用户认知方法，并且在真实的数据集上对提出的方法的准确性和可行性进行了研究。实验结果表明，使用改进的 PageRank 算法和 HITS 算法能够有效地挖掘出知乎平台中具有较为突出特性的关键意见领袖，并且算法的收敛速度较快，具有可复用性和迁移性。通过对关键话题的关键意见领袖进行分析，可以得到有较大潜力的商业化价值。

2　文献综述

目前，在识别关键意见领袖和用户影响力领域，常用的方法包括层次分析法和社交网络挖掘的方式两类。层次分析法是利用关键性的评价指标对关键意见领袖的特征进行描述，建立相应的多层指标体系并进行打分和排序。例如在豆瓣网的意见领袖识别中，利用其中心性、活跃性、吸聚力、传染力等特征指标对用户进行评估[12]；在微博用户影响力评估中，大量的研究都用到了用户的粉丝数、微博数、被转发数、被评论数、知名度、互动率、原创率等指标建立层次分析结构，并由此结构加权平均得到最终的结果[13]。利用层次分析法，可以根据大量的主观研究经验确定较为综合的评价体系，进而得到一个结构性的多目标系统评价模型。

而社交网络挖掘方法通常根据用户的社交网络关系或者转发问答关系来构建相对应的社交网络，再利用基于网络结构的算法或者网络指标来分析用户的重要性，并对最终的结果进行排名。例如利用 PageRank 算法挖掘 Twitter 社交网络中的重要用户[14]；使用 HITS 算法识别微博转发网络中的意见领袖[15]；通过混合评价模型 SALSA 或者加权评价算法得到更为精确的重要性排序结果[16]。基于社交网络的关键意见领袖算法能够挖掘出大规模数据下隐含的规律性知识，进行得到网络信息传播的趋势，由此可以识别出对于整个网络有着关键性影响的用户及其社交网络特性。

利用层次分析法判断社交网络中的用户特性具有更为全面的特性，但是指标的构建在很大程度上需要决策者主观确定，虽然能够在某些角度上体现决策者的知识经验，但也会由于个人的偏好使得构建的系统存在主观随意性。而相比较而言，基于社会网络结构上的关键意见领袖挖掘在客观性和准确性方面具有更大的优势。因此，笔者采用两种方法相结合的方式，利用层次分析法评估用户的活跃度、可信度以及影响力，并用以上用户评估模型的结果为社交网络算法提供方向和依据。

目前较为常用的社交网络挖掘算法为 PageRank 算法，原始的 PageRank 模型是 Google 针对网站中网页重要性评估提出来的。近来大量的研究都致力于对传统的 PageRank 算法进行改进用于挖掘社交网络中的关键用户[17]，并且 PageRank 算法也体现了其独有的高效率和结构稳定性。但是在知乎问答平台中，用户的多次提问与回答的行为是用户的活跃度和信息扩散能力的体现，并不会因为该行为导致影响力的贡献分散。因此，考虑同时使用另外一种常见的网络结构挖掘算法 HITS，用于多角度的度量用户在知乎平台中的社交以及问答行为，并对最终的排序结果进行加权综合，进而形成知乎用户影响力的动态挖掘算法。HITS 算法最初由 Kleinberg 提出，是一种应用于网页排序的结构挖掘算法[18]。在本文中，HITS 算法利用一种相互增强的关系，用于识别满足需求的关键用户和信息传播者。

因此，笔者利用社会网络挖掘的方法识别知乎平台中的关键意见领袖。首先，根据数据集提供的用户关系建立社交网络和问答网络，其次，在该网络结构中利用改进的 PageRank 算法和 HITS 算法动态挖掘用户的影响力。进而，利用排序算法得到用户在各种指标下的影响力排名，最终加权平均得到最终的影响力排序。在此基础上，构建基于话题的关键意见领袖挖掘模型，分析关键意见领袖在网络中的特征的作用，以验证算法的有效性。

3 数据集与网络构建

3.1 数据集

本文所使用的知乎数据集包括 2.6 万知乎平台注册用户的基本信息，即用户的关注数、粉丝数、获得赞同数、获得感谢数、获得收藏数、回答数、提问数、文章数等相关用户特征数据。此外，还获得了其相互之间相对应的关注关系和问答关系。选取这 2.6 万用户中两个较为活跃子集进行分析：获得感谢数大于 1 万的用户（共 1 607 人）、获得感谢数大于 5 万的用户（共 398 人），并将其命名为 Net10K 和 Net50K，其具体的网络结构特性如表 1 所示：

表 1 知乎数据集 Net10K 和 Net50K 具体网络特性描

度量指标	Net10K	Net50K
用户（结点数）	1 607	398
社交网络（社交关系）	120 360	20 310
问答网络（问答关系）	27 846	3 921
强连通分支数量	52	9
平均最短路径长度	2.31	1.86

3.2 社交网络和问答网络的构建

笔者基于知乎社区的用户和内容结构构造了一个双层网络结构（见图 1）。用户与用户之间的相互关注关系构成了第一层网络，例如 U2 与 U3 相互关注，因此 U2 与 U3 之间可以用双向箭头进行连接；U1 关注了 U3，因此从 U1 到 U3 有一条指向 U3 的单向直线。这也是传统的层次分析法的基本思路，根据用户的静态属性，即粉丝数、微博数、转发数等来进行判别。问题回答网络由用户提出的问题和用户的回答构成，例如在图 1 中用户 U2 和 U5 分别提出了问题 Q1 和 Q2，问题 A1 和 A2 回答了 Q1，因此从 A1 和 A2 分别有一条指向 Q5 的单向直线。接下来针对知乎双层网络结构进行抽象建模。

根据知乎网络的社交关系，可以将其定义为一个无权有向图 $G_{SN}=(V,E)$，其中 G_{SN} 表示为知乎社交网络构建的无权有向图；V 表示为顶点集，每一个独立的用户 $v_i \in V$；E 表示为边集，如果用户 v_i 关注了用户 v_i，则 $<v_i,v_j> \in E$；该网络的结构见图 3（a）。而知乎网络对于一个特定话题的问答关系，可以被定义为一个加权有向图 $G_{OA}=(V,E,W,P)$，其中 G_{OA} 表示为知乎问答网络构建的加权有向图；V 表示为顶点集，而每一个顶点代表一个独立的知乎用户；E 表示为边集，如果 $<v_i,v_j> \in E$ 则表示用户 v_i 回答了用户 v_j 提出的一个问题；W 表示为边权重向量，w_{ij} 表示为用户 v_i 和用户 v_j 之间问答关系的数量；P 表示为顶点的强度，由于在真实的网络中每个用户具有不同的影响力，因此根据层次分析法确定的用户综合评价模型来定义 $p(i)$；该网络的结构见图 3（b）。

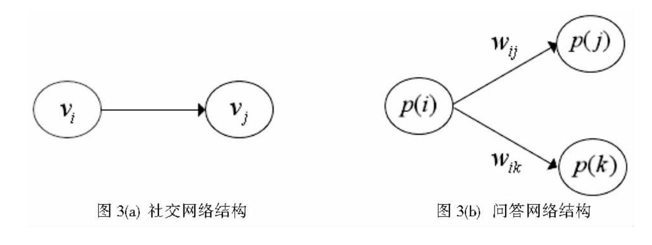


图 3 知乎网络结构

4 关键意见领袖挖掘模型

4.1 层次分析法构建评价指标模型

层次分析法是一种利用构建评价指标的层级结构，是一种定量和定性分析的决策方法。在评价标准中，笔者根据知乎数据集的特征设定了 3 个一级指标以及 8 个二级指标，如表 2 所示：

郭博，许昊迪，雷水旺. 知乎平台用户影响力分析与关键意见领袖挖掘[J]. 图书情报工作,2018,62(20):122-132.

表 2　层次分析法评价指标

一级指标	二级指标
活跃度 A	回答数 A_1
	提问数 A_2
	文章数 A_3
可信度 Q	获得赞同数 V
	获得感谢数 T
	获得收藏数 P
影响力 I	粉丝数 F_1
	关注数 F_2

假设在相同一级指标下的二级指标的权重相同，因此各一级指标的公式定义如下：

$$A=\frac{1}{3}(A_1+A_2+A_3),\ Q=\frac{1}{3}(V+T+P)$$

$$I=\frac{1}{2}(F_1+F_2) \qquad 公式(1)$$

其中，各二级指标是由相对应的真实数据归一化得到，归一化公式如下所示：

$$x_0=\frac{x-x_{min}}{x_{max}-x_{min}} \qquad 公式(2)$$

因此，根据层次分析法对于用户 i 的综合评价模型为：

$$U(i)=w_A(i)+w_QQ(i)+w_II(i) \qquad 公式(3)$$

其中，活跃度 A、可信度 Q、影响力 I 权值的矩阵向量为 $w=(w_A,w_Q,w_I)$，该算法如表 3 所示：

表 3　层次分析法算法概述

```
层次分析法：
Initialize：N：G 中总顶点数
A(i)：用户 i 的活跃度度量
Q(i)：用户 i 的可信度度量
I(i)：用户 i 的影响力度量
各指标的权值的矩阵向量为 w = (w_A, w_Q, W_I)
i = 1
While (i≤N)
    U(i) = w_A A(i) + w_Q Q(i) + w_I I(i)
    i = i + 1
return
    Ui)：每个用户 i 的综合评价模型指标
```

4.2　基于 PageRank 算法关键意见领袖挖掘模型

PageRank 算法是数据挖掘领域较为常见的一种算法，该算法利用每个页面的权威值来评估网站中网页的重要性。页面的权威值被定义为指向该页面的其他页面平均分配给该页面的权威值之和，通过迭代计算可以得到最终的网页等级划分[19]。假设用户在社交网络与问答网络中和其他的用户都具有相对应的交互关系，笔者将知乎排名值(Zhihu-Rank)和 PageRank 算法相结合，可以得到如下改进的 PageRank 算法，用于识别知乎平台中的意见领袖和关键用户。

由于社交网络是一个无权有向网络，根据传统的 PageRank 算法将顶点 的社交网络排名 定义如下：

$$SR(i)=(1-\alpha)+\alpha\sum_{j\to i}SR(j)/d_j \qquad 公式(4)$$

其中，d_i 是结点 j 的出度，α 是阻尼系数，其取值范围为在 0 到 1 之间，在大多数情况下将 α 取为0.85[20]。在真实的网络中，如果一个用户具有大量的粉丝而没有关注其他用户，就可能造成在该顶点出现权威值滞留的现象，进而使得传递受到阻碍。因此，引入了随机冲浪模型以及阻尼系数 α 用来进行随机跳转，解决权威值不断滞留的现象[21]。计算社交网络各用户影响力的 PageRank 算法见表 4。在该算法中为了得到更为精确的结果，设定用于控制迭代结束的参数 e 为 10^{-20}。

表 4　社交网络 PageRank 算法概述

```
基于 PageRank 算法挖掘社交网络中关键意见领袖：
Initialize：
N：G 中总顶点数
e：完全收敛达到的精度为 10^-20
SR_0(i)：问答网络初始权重，SR_0(i) = 1/N
While( ∑_{i=1}^{N} |SR_t(i) - SR_{t-1}(i)|^2 ≥ e)
    SR(i) = (1 - α) + α∑(SR(j)/d_j)
return
    SR(i)：每个用户 i 在社交网络中得到的 PageRank 值
```

知乎平台中的问答网络是一个加权有向的网络，因此定义两个顶点之间的边权重如下：

$$w_{ij}=p(i)\cdot N_{ij} \qquad 公式(5)$$

其中，$p(i)$ 为层次分析法根据每个用户的活跃度、可信度以及影响力得到的用户影响力综合指标，N_{ij} 是用户 i 和用户 j 在问答关系中出现的次数。由于问答网络需要考虑权重，因此在计算的时候需要在每个顶点形成权威值的不对等传递，以真实反映每个用户的影响力。每个顶点 i 在问答网络 $SR(i)=(1-\alpha)+\alpha\sum_{j\to i}SR(j)/d_j$ 的排名 $QR(i)$ 可以被定义为：

$$QR(i)=(1-\alpha)+\alpha\sum_i(QR(j)w_{ji}/\sum_k w_{jk}) \qquad 公式(6)$$

根据对应的问答网络和社交网络计算出的知乎排名，可以得到用户 i 在整个知乎网络中的综合排名，计算公式如下所示：

$$ZR(i)=w_jSR(i)+w_2QR(i) \qquad 公式(7)$$

其中，$w=(w_1,w_2)$ 为问答网络和社交网络的权重向量。当得到所有用户的知乎排名值之后，将其排序取前 N 个用户为最终在两个网络中的关键意见领袖。

计算问答网络各用户影响力的 PageRank 算法如表 5 所示：

表 5 问答网络 PageRank 算法概述

基于 PageRank 算法挖掘问答网络中关键意见领袖：
Initialize：
N:G 中总顶点数
e:完全收敛达到的精度为 10^{-20}
$p(i)$:从层次分析法用户综合评价模型中得到每个用户的活跃度
N_{ij}:用户 i 和用户 j 在问答关系中出现的次数
w_{ij}:用户 i 和用户 j 之间边的权重
$QR_0(i)$:问答网络初始权重,$QR_0(i)=1/N$
$While(\sum_{i=1}^{N}\lvert QR_t(i)-QR_{t-1}(i)\rvert^2\geqslant e)$
$QR(i)=(1-\alpha)+\alpha\sum(QR(j)w_{ji}/\sum w_{jk})$
return
$QR(i)$:每个用户 i 在问答网络中得到的 PageRank 值

4.3 基于 HITS 算法关键意见领袖挖掘模型

最初 HITS 算法也应用于网页排序,其本质是一种网络图的结构挖掘。HITS 算法将网页分成中心网页和权威网页两大类。权威网页被定义为普遍认可的对于特定主旨的重要网页,而中心网页被定义为指向多个与特定主旨相关的权威网页的网页。权威网页和中心网页之间存在相互加强的依赖关系,一个好的权威网页应该被很多好的中心网页所指向;而一个好的中心网页应该指向很多好的权威网页[22]。HIFS 算法利用了权威网页和中心网页之间互相加强的关系,通过迭代计算得到网页排序结果。对应到知乎的社交网络和问答网络中,权威值可以被认为意见领袖的体现,而中心值可以被认为是信息传播者的体现。

因为,笔者将知乎的社交网络定义为无权有向网络,根据传统的 HITS 算法,可以将每一次迭代步骤 t 定义如下:

$$a_i^{(t+1)}=\sum_{j\to i}h_j^{(t)}\qquad h_i^{(t+1)}=\sum_{i\to j}a_j^{(t)}\qquad \text{公式(8)}$$

其中,a_i 是用户 i 的权威值,h_i 是用户 i 的中心值,通过不断迭代直到拟合得到在社交网络下用户 i 的最终权威值和中心值,设定用于控制迭代结束的参数 e 为 10^{-20}。计算社交网络各用户影响力的 HITS 算法见表 6。

知乎平台中的问答网络是一个加权有向的网络,因此可以定义权重如公式(5),将定义中的权重考虑进问答模型的 HITS 算法中:

在计算权威值的过程中考虑用户的综合影响力,而计算中心值的过程中为了避免过分放大层次分析法模型,因此不考虑加入权值进行迭代。根据以上提出的 HITS 传统算法,对计算中心值和权威值迭代公式进行改进,得到如下公式:

表 6 社交网络 HITS 算法概述

基于 HITS 算法挖掘社交网络中关键意见领袖：
Initialize：
N:G 中总顶点数
e:完全拟合达到的精度为 10^{-20}
$a_0(i)$:社交网络的权威值初始化为 $a_0(i)=1/\sqrt{n}$
$h_i(i)$:社交网络的中心值初始化为 $h_0(i)=1/\sqrt{n}$
t:迭代次数 $t=1$
$While(\sum_{i=1}^{N}\lvert a_i^{(t)}-a_i^{(t-1)}\rvert^2+\lvert h_i^{(t)}-h_i^{(t-1)}\rvert^2\leqslant e)$
$a_i^{(t+1)}=\sum h_j^{(t)}\qquad h_i^{(t+1)}=\sum a_j^{(t)}$
$\sum_i(a_i^{(t+1)})^2=1\qquad \sum_i(h_i^{(t+1)})^2=1$
return
$AUTH_SR(i)=a_i$(社交网络中每个用户在 HITS 算法下的权威值)
$HUB_SR(i)=h_i$(社交网络中每个用户在 HITS 算法下的中心值)

$$a_i^{(t+1)}=\sum_{j\to i}h_j^{(t)}\cdot w_{ij}\qquad h_i^{(t+1)}=\sum_{i\to j}a_j^{(t)}\qquad \text{公式(9)}$$

由于在 HITS 算法中考虑到了权重以及问答关系,将该算法到达拟合条件中得到的 a_i 作为 $AUTN_QA(i)$,h_i 作为 $HUB_QA(i)$。

根据对应的问答网络和社交网络计算出的 HITS 知乎排名,可以得到用户 i 在整个知乎网络中的权威值和中心值的综合排名,计算公式如下所示:

$$ZR_AUTH(i)=w_1AUTH_SR(i)+w_2AUTH_QA(i)\qquad \text{公式(10)}$$

$$ZR_HUB(i)=w_1HUB_SR(i)+w_2HUB_QA(i)\qquad \text{公式(11)}$$

5 实验与分析

5.1 网络结构

知乎的社交网络结构由大规模用户的相互关注关系构成,而问答网络结构由用户之间相对应的问答关系构成。由于计算量的限制,笔者选取了 Net10K 和 Net50K 两个活跃的子数据集来挖掘知乎社交网络和问答网络中 TOP10 关键意见领袖。为了分析网络的基本特性,在数据集 Net10K 上绘制入度分布图。图 4(a)和(b)的入度分布图显示了知乎社交网络和问答网络都是一个无标度网络,因为它们都服从幂率分布。

由图 4(a)可以看出,基于社交网络的度分布大量集中在 500 – 2 000 之间,而入度大于 7 000 的用户不足 0.1%。而基于问答网络的度分布大量集中在 100 – 250 之间,此网络的幂指数 $\gamma=1.78$。这说明在问答网络中存在一个较为稀疏的邻接矩阵,但是网络中入度比较大的结点较少,因此说明权威用户在问答网络中的分布较为分散。进一步统计后发现,在问答网络

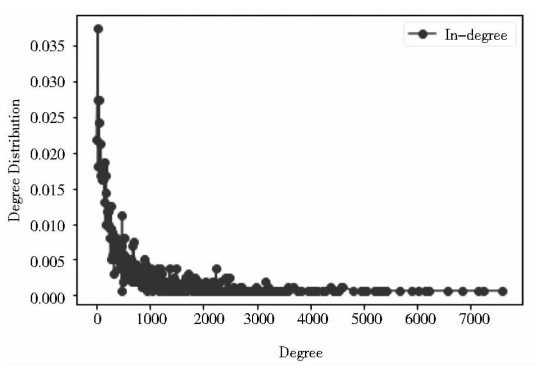


图 4(a)　知乎社交网络度分布

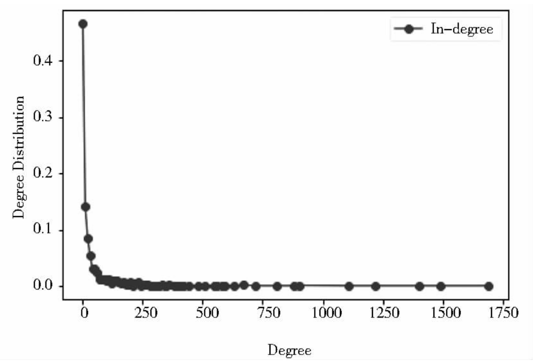


图 4(b)　知乎问答网络度分布

所有的 27 846 条问答关系中，前 100 名回答者所占比例为 49%，而前 100 名问题者仅占比例为 2.8%。

其中，对 Net10K 社交网络中的用户进行社交网络中多种维度上的评估，得到该用户集的基本统计情况，如表 7 所示：

表 7　Net10K 数据集基本统计情况 TOP10 活跃用户

名称	赞同数排名	入度排名	特征向量中心度排名	介数中心度排名	近性中心度排名	综合排名
张佳玮	1	1	4	20	4	1
朱炫	7	10	5	24	6	2
Yolfilm	12	5	20	17	15	3
马伯庸	25	17	3	30	2	4
黄继新	70	3	1	1	3	5
肥肥猫	5	28	24	11	11	6
梁边妖	55	15	4	4	5	7
vczh	4	13	40	8	18	8
寺主人	10	16	20	7	30	9
谢熊猫君	15	22	11	28	9	10

从表 7 可以看出，基于 Net10K 数据集笔者从赞同数、入度、特征向量中心度、介数中心度、近性中心度 5 个角度对数据集中的用户进行了评估，并得到了最终具有 TOP10 影响力的用户。其中，赞同数代表该用户被其他用户赞同的总数目，在知乎中每一次点赞可以代表该用户在知识、经验分享的过程中获得了理解和赞同，是权威度和曝光度的一种展现；入度代表的是该用户被关注的数目，在知乎中被关注的数目越多，代表该用户具有较大的凝聚力和重要性；特征向量中心度代表的是一个用户的重要性取决于其邻居结点的数量，也取决于其邻居结点的重要性，因此一个用户的特征向量中心度越高，代表被越多重要的用户关注；介数中心度代表经过某个点的最短路径数目。一个用户的介数中心度越高，说明其他点之间的最短路径很多甚至全部都必须要经过它中转，因此它成为了很多用户之间交流的通道；近性中心度代表一个结点和其他结点之间的接近程度。一个用户的近似中心度越高，代表该用户与网络中其他用户的距离总体来说较近，反之则较远。利用以上 5 个指标，可以从各个角度衡量每个用户在社交网络中的影响力，得到一个基于网络特性的综合 TOP10 影响力排名。

5.2　基于 PageRank 算法模型结果与分析

根据改进的 PageRank 算法，表 8 和表 9 给出了从 Net10K 和 Net50K 数据集中的社交网络和问答网络中分别得到的 PageRank 值(分别为 SR 值和 QR 值)，并对其进行加权得到最终的综合知乎排名值 Zhihu - Rank。表 8 和表 9 中显示了知乎社交网络和问答网络 Net10 和 Net50 中具有前 10 名 Zhihu-Rank 值的关键用户。

表 8　基于 PageRank 算法模型 Net10K：关键意见领袖 TOP10

排名	名称	SR 值	QR 值	Zhihu-Rank
1	张家玮	0.003 87	0.007 12	0.005 49
2	黄继新	0.004 70	0.003 20	0.003 95
3	张亮	0.003 25	0.004 15	0.003 70
4	马伯庸	0.003 22	0.003 12	0.003 17
5	yolfilm	0.002 77	0.003 50	0.003 14
6	Raymond Wang	0.002 72	0.00342	0.003 07
7	周源	0.003 12	0.002 85	0.002 99
8	葛巾	0.002 50	0.002 69	0.002 60
9	张小北	0.002 65	0.002 52	0.002 58
10	梁边妖	0.002 70	0.002 45	0.002 57

由表 8 看出，基于 Net10K 数据集挖掘得到的关键用户 TOP10 和用户的基础统计集 TOP10 具有较多的重合用户，这说明通过 PageRank 算法得到的关键意见领袖具有有效性，可以较为客观地反映出知乎平台中的具有突出特性、并且具有信息引导和传播作用的关键用户。

表 9 基于 PageRank 算法模型 Net50K：关键意见领袖 TOP10

排名	名称	SR 值	QR 值	Zhihu-Rank
1	张佳玮	0.012 5	0.043 2	0.027 9
2	yolfilm	0.010 2	0.014 8	0.012 5
3	vczh	0.009 2	0.013 9	0.011 5
4	Raymond Wang	0.008 5	0.013 5	0.011 0
5	朱炫	0.009 5	0.012 1	0.010 8
6	cOMMANDO	0.008 2	0.012 0	0.010 1
7	马伯庸	0.008 9	0.010 8	0.009 9
8	梁边妖	0.009 5	0.009 2	0.009 3
9	王冢	0.007 6	0.011 0	0.009 3
10	葛巾	0.009 3	0.008 9	0.009 1

由于 Net10K 和 Net50K 的结点个数具有较大的差异，而网络特性中由于 Net50K 和 Net10K 的平均最短路径程度分别为 2.31 和 1.86，由此可以说明 Net50K 比 Net10K 的网络关注圈更为紧密。因此，基于这两个数据集构建的 PageRank 算法能够挖掘出在不同的网络紧密度（网络大小不同）以及多种网络架构（社交网络和问答网络下）的关键意见领袖，表 8 和表 9 分别反映了在网络大小和紧密度不同的条件下得到的 TOP10 关键意见领袖。由表 8 和表 9 可以看出，其中部分关键意见领袖在两个不同的数据集中都得到了较大的 PageRank 值，这说明该领袖的权威性和影响力不受其外界社交网络范围的缩小而改变，并且其社交圈和问答圈也具有较强的稳定性和强影响力。

根据 PageRank 算法的结果，可以将 Net10K 中所有用户的 zhihu-Rank 进行排名。我们选取了 PageRank 模型下的前 40 个关键意见领袖，对其关注数、关注者数、回答数、提问数、以及可信度进行度量，并将结果绘制在图 5 中。

从图 5 中可以看出：当 N 大于 25 时，能够决定一个用户影响力的多种参数指标将都会表现的非常稳定；而当 N 小于 10 时，多种参数指标就会产生较大程度的波动。因此，对于 TOP N 关键意见领袖的关注范围应该在 10 – 20 之间，可以找到较为准确和客观的关键用户。因此，在表 5 中选择 TOP10 用户为合理的研究区间范围。与此同时，从图 5 中可以发现关注的 TOP10 领袖都具有较多的回答数和较多的粉丝数。在社交网络中，更多的粉丝以及在问答网络中更多有意义的回复都意味着在网络中获得更强烈的关注，由此可以从侧面证明笔者的分析是合理的。

5.3 基于 HITS 算法模型结果与分析

根据改进的 HITS 算法模型，表 10 和表 11 分别给

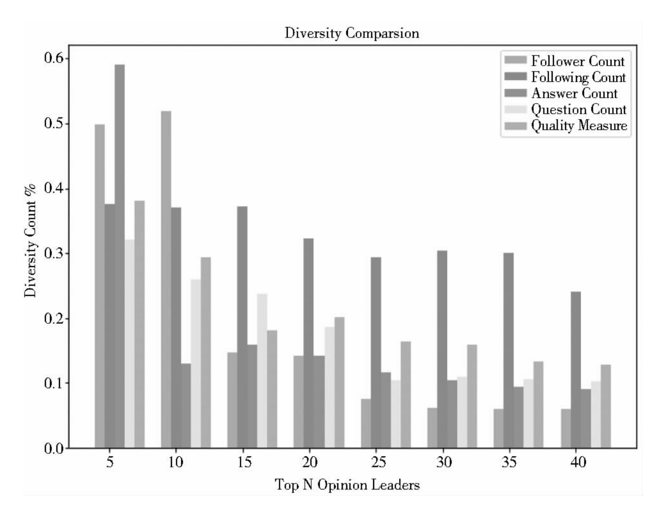


图 5 TOP N 关键意见领袖信息的多样性比较

出了从 Net10K 和 Net50K 数据集中的社交网络和问答网络中分别得到的权威值（分别为 AUTH_SR 和 AUTH_QR）以及中心值（分别为 HUB_SR 和 HUB_QR），并对其加权得到了最终的知乎权威排名值 ZR_AUTH 和知乎中心排名值 ZR_HUB。表 10 和表 11 中分别显示了知乎社交网络和问答网络 Net10 和 Net50 中具有前 5 名 ZR_AUTH 和 ZR_HUB 值的关键用户。

表 10(a) 基于 HITS 算法模型 Net10K：权威值 TOP5

排名	名称	AUTH_SR 值	AUTH_QR 值	ZR_AUTH
1	张佳玮	0.005 28	0.006 12	0.005 70
2	黄继新	0.004 92	0.004 39	0.004 66
3	梁边妖	0.005 34	0.003 82	0.004 58
4	马伯庸	0.005 10	0.004 02	0.004 56
5	葛巾	0.004 76	0.003 76	0.004 26

表 10(b) 基于 HITS 算法模型 Net10K：中心值 TOP5

排名	名称	HUB_SR 值	HUB_QR 值	ZR_HUB
1	yolfilm	0.004 59	0.004 28	0.004 35
2	杨大懒人	0.005 14	0.003 12	0.004 13
3	徐湘楠	0.004 28	0.003 99	0.004 13
4	Edison Chen	0.003 74	0.002 88	0.003 31
5	ZENHO	0.003 67	0.002 54	0.003 11

表 11(a) 基于 HITS 算法模型 Net50K：权威值 TOP5

排名	名称	AUTH_SR 值	AUTH_QR 值	ZR_AUTH
1	张家玮	0.007 25	0.006 21	0.006 73
2	梁边妖	0.007 18	0.003 12	0.005 15
3	朱炫	0.005 68	0.004 61	0.005 15
4	马伯庸	0.005 60	0.004 52	0.005 06
5	谢熊猫君	0.004 91	0.004 98	0.004 95

表 11(b)　基于 HITS 算法模型 Net50K:中心值 TOP5

排名	名称	HUB_SR 值	HUB_QR 值	ZR_HUB
1	yolfilm	0.006 28	0.005 82	0.006 05
2	采铜	0.005 26	0.005 23	0.005 25
3	寺主人	0.005 10	0.004 26	0.004 68
4	肥肥猫	0.004 98	0.004 11	0.004 55
5	Raymond Wang	0.00388	0.00512	0.004 50

根据表 10(a)的结果,可以看出在 AUTH 方面(具有较高权威值)的 TOP5 用户全部都出现在 PageRank 算法获得的关键意见领袖 TOP10 中,说明在挖掘知乎平台的社交网络和问答网络中的关键意见领袖时,PageRank 值和 HITS 的权威值都较好地展现了一个用户在两个网络中的综合影响力和权威性。而在表 10(b)中,可以看出推荐的用户大多数都不属于 PageRank 算法获得的关键意见领袖 TOP10 集合中,甚至有一些用户在基本统计集 TOP10 中也没有出现过。由此可以表明,在 HUB 方面(具有较高中心值)的 TOP5 用户虽然没有在社交网络或问答网络中表现出较高的 Zhihu-Rank 值,但都具有较高的介数中心值和近似中心值,进而具有较强的信息传播和交流能力。

图 6(a) 权威值 TOP5 的各指标

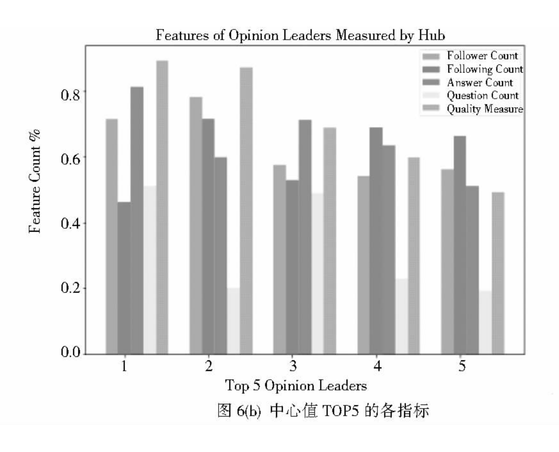


图 6(b) 中心值 TOP5 的各指标

图 6　基于 Net10 网络中的 TOP5 关键意见领袖相关数据

图 6(a)和图 6(b)分别反映了 HITS 算法从 Net10 网络中挖掘出的 TOP5 关键意见领袖在基本统计集上面的指标。由图 6(a)可以总结出,具有较高权威值的用户特征一般在于单一特征的绝对优势,例如具有较高的关注度、回答数、以及用户评估模型中较高的影响力。由图 6(b)可以总结出,具有较高中心值的用户特征一般在于各种指标水平都相对平均,从各个角度都存在较强的信息传播的能力。

综合来看,HITS 算法和 PageRank 算法挖掘出的关键意见领袖有大量的重合用户,因为 PageRank 算法获得的关键意见领袖排序值大致近似于 HITS 算法中 AUTH(权威值)和 HUB(中心值)的一种有序叠加。例如用户张亮的 AUTH 值和 HUB 值在 Net10K 网络中都不在 TOP5 中,但都排在 TOP15 之间,而 PageRank 值却排在 TOP3,因此可以证明很多用户既能受到广大关注者的吸引,又是信息传播的渠道。

5.4　算法评估

图 7 显示了 HITS 算法和 PageRank 算法在关键意见领袖识别过程中的拟合率。由图 7 可以看出,知乎数据集在 HITS 算法和 PageRank 算法拟合的过程中收敛的速度都相对较快,并且 PageRank 算法比 HITS 算法得到了更好的拟合率。而 PageRank 算法的拟合曲线在训练样本数量的 20% 前持续上升,而在 20% 之后就达到了一个稳定的状态。相比较而言,HITS 算法相对来说收敛速度较慢,需要更多的训练样本来进行迭代。

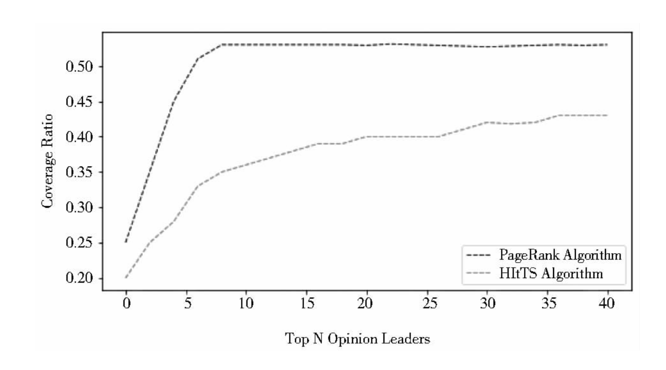


图 7　PageRank 算法和 HITS 算法拟合比较

根据 PageRank 算法得到的 Zhihu-Rank 值,HITS 算法得到的 AUTH 值、HUB 值,以及用户本身的赞同数、粉丝数、特征向量中心度、介数中心度、近性中心度从不同角度反映用户的影响力。因此,进一步对用户

的 PageRank 值、AUTH 值、HUB 值以及 5 个评价指标进行相关性分析,以研究其内在的联系,用于验证算法的有效性。

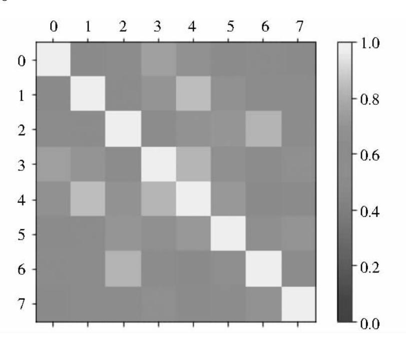


图 8 各指标相关性分析

其中,图 8 中的 0,1,2,3,4,5,6,7 分别代表以上的 8 个指标:PageRank 值、AUTH 值、HUB 值、赞同数、粉丝数、特征向量中心度、介数中心度以及近似中心度。

通过图 8 可以看出,Zhihu-Rank 和 HITS 算法得到的权威值 AUTH、特征向量中心度、近似中心度存在中等的相关性,这说明 Zhihu-Rank 值与 AUTH、特征向量中心度、近似中心度对于用户影响力的衡量并不相同,是几个具有不同含义的度量值。而 Zhihu-Rank 与赞同数和粉丝数具有较强相关性。考虑到一个粉丝数与赞同数较高的用户被关注及做出回答的概率较大,因此可以证明该用户在 PageRank 算法里较为重要的结果是合理的。

而从权威值 AUTH 与中心值 HUB 的相关性值可以看出,虽然 HITS 算法是利用社交网络的关系建立转移矩阵用来挖掘用户重要性,但仍然较好地体现了用户的基本统计属性。例如其粉丝数、赞同数、以及各类网络指标都在其中得到了有效平衡。

特别的,虽然权威值 AUTH 和中心值 HUB 存在中等相关性,但中心值和权威值在 TOP20 的排名上均出现了较大程度上的重合,这说明挖掘出的关键意见领袖不仅是具有强大影响力的人,也是信息的传播途径。

5.5 特定话题下关键意见领袖分析

为了验证算法的有效性,笔者通过研究某科技公司的话题“XX 科技”,来检验该话题客观与准确性。此处考虑论文的客观与中立性,避免软广嫌疑,作者隐去了具体公司名称。。因此,我们利用 PageRank 算法和 HITS 算法分别挖掘出在“XX 科技”下的关键话题依赖用户,其实验结果如表 12 和表 13 所示:

表 12 基于 PageRank 算法模型“XX 科技”:关键意见领袖 TOP5

排名	名称	SR 值	QR 值	Zhihu-Rank
1	李楠	0.009 28	0.007 28	0.008 28
2	贾敬贤	0.006 21	0.007 08	0.006 65
3	黄梁一觉	0.006 10	0.006 89	0.006 50
4	Wong Xu	0.005 10	0.006 92	0.006 01
5	LJRFOX STUDIO	0.005 02	0.006 99	0.006 01

表 13(a) 基于 HITS 算法模型“XX 科技”:权威值 TOP5

排名	名称	HUB_SR 值	HUB_QR 值	ZR_HUB
1	李楠	0.006 19	0.005 21	0.005 70
2	胡杰	0.005 50	0.003 10	0.004 30
3	刘海光	0.004 92	0.003 10	0.004 01
4	黄轶轩	0.005 23	0.002 62	0.003 92
5	魅蓝科技	0.005 71	0.002 10	0.003 91

表 13(b) 基于 HITS 算法模型“XX 科技”:中心值 TOP5

排名	名称	AUTH_SR 值	AUTH_QR 值	ZR_AUTH
1	吴章金 falcon	0.005 11	0.005 00	0.005 05
2	楊甚麼	0.005 22	0.004 38	0.004 80
3	邓邓大人	0.004 82	0.003 90	0.004 36
4	黄梁一觉	0.005 02	0.003 53	0.004 28
5	Wong Xu	0.004 99	0.003 38	0.004 19

表 14 知乎认证“XX 科技下”优秀回答者

排名	名称	粉丝数	赞同数	感谢数
1	李楠	419 083	204 346	35 897
2	胡杰	18 894	121 780	23 262
3	邓邓大人	3 905	15 989	3 348

由表 12 和表 13 可以看出基于话题“XX 科技”下的 TOP5 关键意见领袖,虽然 PageRank 算法和 HITS 算法从不同角度对用户的影响力进行了衡量,但二者 TOP5 的结果中仍然出现了重合。例如知乎用户李楠、黄梁一觉和 Wong Xu 在 PageRank 算法中属于 TOP5,在 HITS 算法中权威值或中心值也出现在 TOP5,这说明该类用户在知乎关键意见领袖的网络中即是“权威用户”,又是“信息传播者”。因此,挖掘这类用户无论是在商业营销中,还是用户影响力学术研究中都会产生巨大的参考价值。

表 14 为知乎平台认证的基于该话题的优秀回答者,其中包括了三位比较的知名的用户李楠、胡杰以及邓邓大人。而从表 12 和表 13 显示的结果可以看出,这三位用户都被笔者提出的算法从不同角度挖掘出

来,成为知乎平台中具有话题依赖的关键意见领袖。由以上的结果也可以看出,笔者提出的算法具有较强的实用性和准确性。

6 结语

以知乎用户之间的社交关系和问答关系构建知乎社交网络和问答网络,通过层次分析法进行指标分层,建立用户影响力评价模型,并应用到后续的 PageRank 算法和 HITS 算法的权重选取上,用于挖掘知乎平台中具有较强影响力的关键领袖。通过结合社交网络和问答网络中知乎用户的重要性指标,克服了单一网络中的衡量关键领袖算法的片面性和局限性。同时,通过改进的 PageRank 算法和 HITS 算法,考虑到用户活跃度、可信度和影响力,用于进一步地提升关键领袖挖掘过程中的准确性和客观性,使得对于关键领袖的影响力分析更为全面。实验结果表明,笔者提出的 PageRank 算法和 HITS 算法能够有效地挖掘出知乎平台中具有较为突出特性的关键领袖,并且算法的收敛速度较快,具有可复用性和迁移性。从话题依赖用户的研究中,其结果显示了挖掘关键话题领袖和影响力的潜在的巨大商业价值,在未来经过不断地修正和完善,预估该模型可以达到驱动和指导营销的作用。文章也存在一些不足,并没有考虑到"水军"[23]群体对计算关键领袖影响力的负面作用,这也是今后笔者研究的另外一个重点。

参考文献:

[1] 贾佳, 宋恩梅, 苏环. 社会化问答平台的答案质量评估——以"知乎","百度知道"为例[J]. 信息资源管理学报, 2013, 3(2): 19－28.

[2] 宋好. 微博时代"意见领袖"特点探析[J]. 今传媒, 2010, 11(41): 1.

[3] 王秀丽. 网络社区意见领袖影响机制研究——以社会化问答社区"知乎"为例[J]. 国际新闻界, 2014, 36(9): 47－57.

[4] 罗晓光, 溪璐路. 基于社会网络分析方法的顾客口碑意见领袖研究[J]. 管理评论, 2012, 24(1): 75－81.

[5] 彭兰. 网络中的人际传播[J]. 国际新闻界, 2001, 3(1):47－53.

[6] 曾润喜. 网络舆情管控工作机制研究[J]. 图书情报工作, 2009, 53(18): 79－82.

[7] 刘志明, 刘鲁. 微博网络舆情中的意见领袖识别及分析[J]. 系统工程, 2011, 29(6): 8－16.

[8] 郭博,赵隽瑞,孙宇. 社会化问答社区用户行为统计特性及其动力学分析: 以知乎网为例[J]. 数据分析与知识发现, 2018, 2(4): 48－58.

[9] 郭博,李守光,王昊,等. 电商评论综合分析系统的设计与实现——情感分析与观点挖掘的研究与应用[J]. 数据分析与知识发现,2017,1(12): 1－9.

[10] YOGANARASIMHAN H. Impact of social network structure on content propagation: a study using YouTube data[J]. Quantitative marketing and economics, 2012, 10(1): 111－150.

[11] ZHAI Z, XU H, JIA P. Identifying opinion leaders in BBS[C]//Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Piscataway: IEEE Computer Society, 2008: 398－401.

[12] 丁汉青, 王亚萍. SNS 网络空间中"意见领袖"特征之分析——以豆瓣网为例[J]. 新闻与传播研究, 2010 (3): 82－91.

[13] 杨长春, 王天允, 叶施仁. 微博意见领袖影响力评价指标体系研究——基于媒介影响力视角[J]. 情报杂志, 2014, 33(8): 178－183.

[14] KWAK H, LEE C, PARK H, et al. What is Twitter, a social network or a news media? [C]//Proceedings of the 19th international conference on World wide web. New York: ACM, 2010.

[15] XIE Y, HUANG T Z. A model based on cocitation for web information retrieval[J]. Mathematical problems in engineering, 2014: 1－6.

[16] LANGVILLE A N, MEYER C D. Google's PageRank and beyond: the science of search engine rankings[M]. Princeton : Princeton University Press, 2011.

[17] 毛国君, 谢松燕, 胡殿军. PageRank 模型的改进及微博用户影响力挖掘算法[J]. 计算机应用与软件, 2017, 34(5): 28－32.

[18] MILLER J C, RAE G, SCHAEFER F, et al. Modifications of Kleinberg's HITS algorithm using matrix exponentiation and web log records[C]//Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. New York : ACM, 2001: 444－445.

[19] LIN Y, LI H, LIU X, et al. Hot topic propagation model and opinion leader identifying model in microblog network[C]//Abstract and Applied Analysis. London: Hindawi Publishing Corporation, 2013.

[20] XING W, GHORBANI A. Weighted pagerank algorithm[C]//Conference on Communication Networks and Services Research.. Piscataway : IEEE, 2004: 305－314.

[21] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the web[R]. Stanford : Stanford InfoLab, 1999.

[22] 熊涛, 何跃. 微博转发网络中意见领袖的识别与分析[J]. 现代图书情报技术, 2013 (6): 55－62.

[23] GUO B, WANG H, YU Z, et al. Detecting spammers in e-commerce website via spectrum features of user relation graph[C]//Advanced Cloud and Big Data (CBD), 2017 Fifth International Conference on. Piscataway :IEEE, 2017: 324－330.

作者贡献说明：

郭博：提出研究思路，设计研究方案，算法编写，论文修改；

许昊迪：完成知乎关键领袖挖掘实验和分析，撰写论文；

雷水旺：论文修改。

Analysis of User Influence and Identification of Key Opinion Leaders Based on Zhihu Platform

Guo Bo[1] Xu Haodi[2] Lei Shuiwang[3]

[1] Meizu Telecom Equipment Co., Ltd. Beijing 100872

[2] The Hong Kong University of Science and Technology, Hong Kong 999077

[3] Xuzhou College of Industrial Technology Library, Xuzhou 221140

Abstract: [**Purpose/significance**] With the rapid development of network technology, the platform of Zhihu has become a significant carrier to discuss social public topics and share knowledge as well as specified experience. Therefore, it is of importance for studying the communication channels of social network information to investigate the influence of key users and dig out the key opinion leaders in the Zhihu platform. [**Method/process**] By the means of improved PageRank and HITS algorithms, this study constructed a model for evaluating user influence based on the social network and question answering network of Zhihu platform, and identified the key users and opinion leaders accurately and objectively. [**Result/conclusion**] The experimental results show that PageRank and HITS algorithms in this paper could effectively extract several key opinion leaders with prominent features in Zhihu platform, the speed of the convergence is fast and with high reusability and mobility. By processing and analyzing the user data set of Zhihu platform, we successfully build a model for evaluating the user influence and mining key opinion leaders. Along with the verification of specified topics, it can be inferred that this model has enormous application value and commercial promotion prospect.

Keywords: Zhihu user influence key opinion leader PageRank algorithm HITS algorithm